Research article

Support Vector Machines for predicting protein structural class

Yu-Dong Cai*[1], Xiao-Jun Liu[2], Xue-biao Xu[3] and Guo-Ping Zhou[4]

Address: [1]Shanghai Research Centre of Biotechnology, Chinese Academy of Sciences, Shanghai, 200233, China, [2]Institute of Cell, Animal and Population Biology University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, U.K, [3]Department of Computing Science, University of Wales, College of Cardiff, Queens Buildings, Newport Road, PO Box 916, Cardiff CF2 3XF, U.K and [4]Department of Structural Biology, Burnham Institute, La Jolla, California 92037, USA

E-mail: Yu-Dong Cai* - y.cai@umist.ac.uk; Xiao-Jun Liu - x.liu@ed.ac.uk; Xue-biao Xu - x.xu@cs.cf.ac.uk; Guo-Ping Zhou - gbzhou@burnham-inst.org

*Corresponding author

Published: 29 June 2001

BMC Bioinformatics 2001, **2**:3

Received: 24 May 2001

Accepted: 29 June 2001

This article is available from: http://www.biomedcentral.com/1471-2105/2/3

Abstract

Background: We apply a new machine learning method, the so-called Support Vector Machine method, to predict the protein structural class. Support Vector Machine method is performed based on the database derived from SCOP, in which protein domains are classified based on known structures and the evolutionary relationships and the principles that govern their 3-D structure.

Results: High rates of both self-consistency and jackknife tests are obtained. The good results indicate that the structural class of a protein is considerably correlated with its amino acid composition.

Conclusions: It is expected that the Support Vector Machine method and the elegant component-coupled method, also named as the covariant discrimination algorithm, if complemented with each other, can provide a powerful computational tool for predicting the structural classes of proteins.

Introduction

The observed results by Muskal and Kim [1] suggested that the structural class of a protein might basically depend on its amino acid composition. Many efforts [2,3,4,5,6,7,8,9,10,11,12,13,14] have been made to predict the structural class of a protein based on its amino acid composition. The physical mechanism about this kind of correlation has been discussed by Bahar et al. [14] and Chou [15]. For a systematic description in this area, see a comprehensive review by Chou and Zhang [16] and an updated review [17]. In this paper, we try to apply Vapnik's Support Vector Machine [18] to approach this problem. In this work. Support Vector Machine was performed based on the data sets constructed by Zhou [19] based on SCOP [20]. In ref.19 the reason why these data sets are more reasonable has also been addressed. As a result, high rates of self-consistency and jackknife test were obtained. This has further confirmed that the structural class of a protein is considerably correlated with its amino acid composition.

Results and Discussion

Success rate of self-consistency of SVMs

In this research, the examination for the self-consistency of the SVM method was tested. The following two data sets from Zhou [19] are used. One consists of 277 domains, of which 70 all-α domains, 61 all-β domains, 81 α/β domains, and 65 α+β domains. The other data set consists of 498 domains, of which 107 are all-α domains, 126 all-β,136 α/β domains, and 129 α+β domains. All the

Table 1: Results of Self-Consistency Test

Dataset	Algorithm	Rate of correct prediction for each class				Overall Rate of Correct Prediction
		all-α	all-β	α/β	α+β.	
277 domains	component coupled	95.7%	93.4%	95.1%	92.3%	94.2%
	neural network	98.6%	93.4%	96.3%	84.6%	93.5%
	SVM	100%	100%	100%	100%	100%
498 domains	component coupled	95.8%	95.2%	94.9%	95.4%	95.8%
	neural network	100%	98.4%	96.3%	84.5%	94.6%
	SVM	100%	100%	100%	100%	100%

rates of correct prediction for the four structural classes of both datasets reach 100%. These rates are "training" accuracy, indicating that after being trained, the SVM model has grasped the complicated relationship between the amino acid composition and protein structure.

Success rate of jackknife test of SVMs

We use jackknife test for cross-validation. The cross-validation by jackknifing is thought the most objective and rigorous way in comparison with sub-sampling test or independent dataset test [16, 21,22]. During the process of jackknife analysis, the datasets are actually open, and a protein will in turn move from each to the other. As a result, the overall rate of correct prediction for the four structural classes of 277 domains (the 1 st set) was 220/277 = 79.4%; while the rates of correct prediction for the four structural classes of 498 domains (the 2nd set) was 464/498 = 93.2%.

Comparison to neural network method and elegant component-coupled algorithm

Zhou [19] applied the elegant component-coupled algorithm developed by Chou et al. [11,12,13] to protein structure class prediction. Later Cai and Zhou [23] applied neural network method to the same problem. The comparison of their results to SVM method is given in Table 1 (for self-consistency test) and Table 2 (for jackknife test).

The comparison should be focused on the jackknife rates (Table 2) because it represents the rate obtained by following a more objective test procedure [21,22]. From Table 2 we can see that the rates of both the SVM and the component-coupled algorithm are higher than those of neural network. Although the rates obtained here by SVM are slightly higher than those by the component-coupled algorithm, it does not mean the predicted results by SVM are always better than those by the component-coupled algorithm. For some cases, the results obtained by the latter might be better than those by the former. Accordingly, it is expected, the SVM method and the component-coupled algorithm, if complemented with each other, will provide a powerful tool for predicting protein structural class.

Conclusion

The current study has further supported, from the approach of SVMs, the conclusion drawn by Chou and his co-workers [11,12,13] and Zhou [19] that if the coupling effect among different amino acid components can be properly taken into account, the prediction quality of protein structural classes can be significantly improved.

Materials and Methods

Support Vector Machine (SVM)

Support Vector Machine (SVM) is one kind of learning machine based on statistical learning theory. The basic idea of applying SVM to pattern classification can be stated briefly as follows. First, map the input vectors into one feature space (possible with a higher dimension), either linearly or non-linearly, which is relevant with the selection of the kernel function. Then, within the feature space from the first step, seek an optimized linear division, i.e. construct a hyperplane which separates two classes(this can be extended to multi-class). SVM training always seeks a global optimized solution and avoids over-fitting, so it has the ability to deal with a large number of features. A complete description to the theory of SVMs for pattern recognition is in Vapnik's book [24]

SVMs have been used in a wide range of problems including drug design [25], image recognition and text classification [26], microarray gene expression data analysis [27], and protein fold recognition [28].

In this paper, we apply Vapnik's Support Vector Machine [18] for the structural classes of proteins. We download the SVMIight, which is an implementation (in C Lan-

Table 2: Results of Jackknife Test

Dataset	Algorithm	Rate of correct prediction for each class				Overall Rate of Correct Prediction
		all-α	all-β	α/β	α+β	
277 domains	component coupled	84.3%	82.0%	81.5%	67.7%	79.1%
	neural network	68.6%	85.2%	86.4%	56.9%	74.7%
	SVM	74.3%	82.0%	87.7%	72.3%	79.4%
498 domains	component coupled	93.5%	88.9%	90.4%	84.5%	89.2%
	neural network	86.0%	96.0%	88.2%	86.0%	89.2%
SVM		88.8%	95.2%	96.3%	91.5%	93.2%

guage) of SVM for the problem of pattern recognition. The optimization algorithm used in SVMIight can be found in [29,30]. The code has been used in text classification, image recognition [26], microarray gene expression data analysis [27] and protein fold recognition [28].

Suppose we are given a set of samples, i.e, a series of input vectors $X_i \in R^d (i = 1,...,N)$

with corresponding labels $y_i \in \{+1,-1\}(i = 1,...,N)$.

Where -1 and +1 are used to stand respectively for the two classes. The goal here is to construct one binary classifier or derive one decision function from the available samples, which has small probability of misclassifying a future sample. Both the basic linear separable case and the most useful linear non-separable case for most real life problems are considered here:

The linear separable case

In this case, there exists a separating hyper plane whose function is $\vec{W} \bullet \vec{X} + b = 0$, which implies:

$y_i(\vec{W} \bullet \vec{x}_i + b) \geq 1, i = 1,...,N$

By minimizing $\frac{1}{2}\left\|\vec{W}\right\|^2$ subject to this constraint, the SVM approach tries to find a unique separating hyperplane. Here $\left\|\vec{W}\right\|^2$ is the Euclidean norm of $\vec{w}$, which maximizes the distance between the hyper plane, i.e. Optimal Separating Hyperplane or OSH [31], and the nearest data points of each class. The classifier is called the largest margin classifier. By introducing Lagrange multipliers α_i, the SVM training procedure amounts to solving a convex QP problem. The solution is a unique globally optimized result can be shown having the following expansion:

$$\vec{W} = \sum_{i=1}^{N} y_i \alpha_i \cdot \vec{x}_i$$

Only if the corresponding $\alpha_i > 0$ these are $\vec{x}_i$ called Support Vectors. When a SVM is trained, the decision function can be written as:

$$f(\vec{x}) = \text{sgn}(\sum_{i=1}^{N} y_i \alpha_i \cdot \vec{x} \bullet \vec{x}_i + b)$$

Where sgn() in the above formula is the given sign function.

The linear non-separable case

(i) "soft margin" technique.

In order to allow for training errors, ref.31 introduced slack variables:

$\xi_i > 0, i = 1, ..., N$

And relaxed separation constraint is given as:

$$y_i(\overline{w} \bullet \vec{x}_i + b) \geq 1 - \xi_i, (i = 1,...,N)$$

And the OSH can be found by minimizing

$$\frac{1}{2}\left\|\overline{w}\right\|^2 + C\sum_{i=1}^{N}\xi_i$$

Where C is a regularization parameter used to decide a trade- off between the training error and the margin.

(ii) "kernel substitution" technique

SVM performs a nonlinear mapping of the input vector $\bar{x}$ from the input space R^d into a higher dimensional

Hilbert space, where the mapping is determined by the kernel function. Then like in case (i), it finds the OSH in the space *H* corresponding to a non-linear boundary in

the input space. Two typical kernel functions are listed below:

$$K(\vec{x}_i, \vec{x}_j) = (\vec{x}_i \bullet \vec{x}_j + 1)^d$$

$$K(\vec{x}_i, \vec{x}_j) = \exp(-r\left\|\vec{x}_i - \vec{x}_j\right\|^2)$$

And the form of the decision function is

$$f(\vec{x}) = \mathrm{sgn}(\sum_{i=1}^{N} y_i \alpha_i \cdot K(\vec{x}, \vec{x}_i) + b)$$

For a given data set, only the kernel function and the regularity parameter C must be selected to specify one SVM.

The Training and Prediction of Protein Structural Class

According to the SCOP database, the protein domains generally fall into one of the following four classes: (1) all-α, (2) all-β, (3) α/β, (4) α+β.

According to its amino acid composition, a protein domain can be represented by a point or a vector in a 20-D space. However, of the 20 amino acid composition components, only 19 are independent due to the normalisation condition [11]. Accordingly, strictly speaking, if based on amino acid composition, a protein should be represented by a point or a vector in a 19-D space rather than 20-D space as defined in a conventional manner. Furthermore, according to Chou's invariance theorem, the final predicted result will remain the same regardless of which one of the 20 components is left out for forming the 19-D space. It is extremely important to realize this, particularly when the calculations involve a covariance matrix such as in the case ofrefs.11-14. For the current study, the amino acid composition was used as the input of the SVM.

The SVM method applies to two-class problems. In this paper, for the four-class problems, we use a simple and effective method: "one-against-others" method [27, 28] to transfer it into two-class problems.

The computations were carried out on a Silicon Graphics IRIS Indigo work station (Elan 4000).

In this research, for the SVM, the width of the Gaussian RBFs is selected as that which minimized an estimate of the VC-dimension. The parameter C that controls the error-margin tradeoff is set at 100. After being trained, the hyperplane output by the SVM was obtained. This indicates that the trained model, i.e. hyperplane output which is including the important information, has the function to identify protein structural classes.

We first test the self-consistency of the method, latterly is to test the method by cross-validation (jackknife test). As a result, the rates of both self-consistency and cross-validation were quite high.

References

1. Muskal SM, Kim SH: **Predicting protein secondary structure content: A tandem neural network approach** *J. Mol. Biol.* 1992, **225**:713-727
2. Chou PY: **Amino Acid composition of four classes of protein,** *in Abstracts of Papers, Part I, Second Chemical Congress of the North AmericanContinent. Las Vegas, Nevada,* 1980
3. Chou PY: **Prediction of protein structural classes from amino acid composition** *In: Prediction of Protein Structure and the Principles ofProtein Conformation, ed. Fasman, G.D., Plenum Press: New York.* 1989549-586
4. Nakashima H, Nishikawa K, Ooi T: **The folding type of a protein is relevant to the amino acid composition** *J. Biochem* 1986, **99**:152-162
5. Klein P, Delisi C: **Prediction of protein structural class from amino acid sequence** *Biopolymers* 1986, **25**:1659-1672
6. Zhang CT, Chou KC: **An optimization approach to predicting protein structural class from amino acid composition** *Protein Science* 1992, **1**:401-408
7. Dubchak I, Holbrook SR, Kim SH: **Predicting protein secondary structure content: A tandem neural network approach** *Proteins: Structure, Function and Genetics.* 1993, **16**:79-91
8. Metfessel BA, Saurugger PN, Connelly DP, Rich ST: **Cross-validation of protein structural class prediction using statistical clustering and neural networks** *Protein Science.* 1993, **2**:1171-1182
9. Rost B, Sander C: **Combining evolutionary information and neural networks to predict protein secondary structure** *Protein: Struc. Func., and Genetics.* 1994, **19**:55-72
10. Chandonia JM, Karplus M: **Neural networks for secondary structure and structural class prediction** *Protein Science.* 1995, **4**:275-285
11. Chou KC: **A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space** *Proteins: Structure, Function and Genetics,* 1995, **21**:319-344
12. Chou KC, Maggiora GM: **Domain structural class prediction** *Proteins Engineering,* 1998, **11**:523-538
13. Chou KC, Liu W, Maggiora GM, Zhang CT: **Prediction and classification of domain structural classes** *Proteins: Structure, Function and Genetics.* 1998, **31**:97-103
14. Bahar I, Atilgan AR, Jemigan RL, Erman B: **Understanding the recognition of protein structural classes by amino acid composition** *Proteins* 1997, **29**:172-185
15. Chou KC: **A key driving force in determination of protein structural classes** *Biochem. Biophys. Res. Commun.* 1999, **264**:216-224
16. Chou KC, Zhang CT: **Prediction of Protein Structural Classes** *Critical Reviews in Biochemistry and Molecular Biology.* 1995, **30**:275-349
17. Chou KC: **Review: Prediction of protein structural classes and subcellular location** *Current Protein and Peptide Science.* 2001, **1**:171-208
18. Vapnik VN: **The Nature of Statistical Learning Theory** *Springer,* 1995
19. Zhou GP: **An Intriguing Controversy over Protein Structural Class Prediction** *Journal of Protein Chemistry.* 1998, **17**:729-738
20. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of protein database for the investigation of sequence and structures** *Jourani of Molecular Biology* 1995, **247**:536-540
21. Cai YD: **Is it a paradox or misinterpretation?** *Proteins: Structure, Function and Genetics.* 2001, **43**:336-338
22. Zhou GP, Assa-Munt N: **Some insights into protein structural class prediction** *Proteins: Structure, Function and Genetics.* 2001, **44**:57-59

23. Cai YD, Zhou GP: **Prediction of protein structural classes by neural network** *Biochimie.* 2000, **82(8)**:783-5
24. Vapnik VN: **Statistical Learning Theory** *Wiley-Interscience, New York,* 1998
25. Robert B, Matthew T, Sean H, Bernard B: **Drug Design by Machine Learning: Support Vector Machine for Pharmaceutical Data Analysis** *Proceedings of the AISB'00 Symposium on Artificial Intelligence in Bioinformatics.* 20001-4
26. Joachims T: **Text Categorization with Support Vector Machines: Learning with Many Relevant Features"** *Proceedings of the European Conference on Machine Learning, Springer,* 1998
27. Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet C, Ares JM, Haussler D: **Knowledge-based Analysis of Microarray Gene Expression Data by using Support Vector Machines** *Proc. Natl. Acad. Sci.* 2000, **97**:262-267
28. Ding CHQ, Dubchak I: **Multi-class Protein Fold Recognition Using Support Vector Machines and Neural Networks** *Bioinformatics* 2001, **4(17)**:349-358
29. Joachims T: **Making large-Scale SVM Learning Practical** *Advances in Kernel Methods -Support Vector Learning, B. Scholkopf and C. Burges and A. Smola (ed.), MIT Press,* 199911
30. Joachims T: **Transductive Inference for Text Classification using Support Vector Machines** *International Conference on Machine Learning (ICML),1996b*
31. Cortes C, Vapnik VN: **Support vector networks** *Machine Learning.* 1995, **20**:273-293